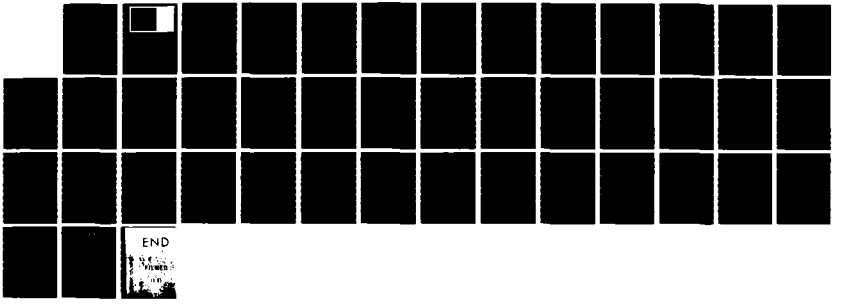
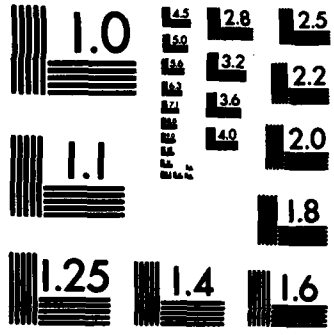
AD-A132 828 ON THE USE OF MARGINAL LIKELIHOOD IN TIME SERIES MODEL ESTIMATION(U) WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER G TUNNICLIFFE-WILSON JUL 83 1/1

UNCLASSIFIED MRC-TSR-2539 DAAG29-80-C-0041 F/G 12/1 NL





MRC Technical Summary Report # 2539

# ON THE USE OF MARGINAL LIKELIHOOD IN TIME SERIES MODEL ESTIMATION

G. Tunnicliffe-Wilson

**Mathematics Research Center**
**University of Wisconsin—Madison**
**610 Walnut Street**
**Madison, Wisconsin 53705**

July 1983

(Received June 3, 1983)





Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

# ON THE USE OF MARGINAL LIKELIHOOD IN TIME SERIES MODEL ESTIMATION

G. Tunnicliffe-Wilson

Technical Summary Report #2539

July 1983

ABSTRACT

This paper is concerned with the estimation of regression models with errors which follow an Autoregressive Integrated Moving Average (ARIMA) process. The effect of the regression upon the ARIMA model parameter estimates is considered and marginal likelihood investigated as a means of overcoming some small sample bias. Examples illustrate the importance of this effect even in samples of moderate size. The consequences regarding inference for the regression coefficients are also discussed.

AMS (MOS) Subject Classifications: 62A10, 62F10, 62M10, 90A20

Key Words: Marginal likelihood, time series estimation, Durbin-Watson test, serial correlation, mixed spectrum

Work Unit Number 4 - Statistics and Probability

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

## SIGNIFICANCE AND EXPLANATION

When simple linear regression is carried out to relate observations of one dependent variable, to observations of one or more independent variables, it is important to allow for any serial correlation in the errors. This helps to avoid nonsense relationships being established, by giving more realistic values for the precision of the estimated coefficients. Unfortunately the regression tends to distort the errors so the evidence of serial correlation may be lost. This paper proposes that modeling of the error correlation should be based only on that information in the observations which cannot be distorted by the regression. Thus it uses what is termed the marginal likelihood criterion. Several examples are used to illustrate how this can improve inference, particularly concerning the question of whether random walk components are present in the error. Applications include models of economic series, and scientific series such as the periods of peak sunspot activity.



The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the author of this report.

# ON THE USE OF MARGINAL LIKELIHOOD IN TIME SERIES MODEL ESTIMATION

G. Tunnicliffe-Wilson

## 1. INTRODUCTION

We shall consider the regression model $y = x\alpha + e$, or in full:

$$y_t = \alpha_1 x_{1,t} + \cdots + \alpha_K x_{K,t} + e_t \quad , \quad t = 1 \cdots n \tag{1.1}$$

where the observations $y$ are dependent upon fixed regressors $x$ which in many cases will be deterministic functions of time $t$.

It is well known that estimation of the coefficients $\alpha$ should take into account the covariance structure of the errors $e$.

Our interest is in situations where $e$ is modelled as either a stationary process with continuous spectrum, or as an integrated stationary process, using the ARIMA (p,d,q) models presented in Box and Jenkins (1976). Thus typically in the stationary case,

$$e_t = \phi_1 e_{t-1} + \cdots + \phi_p e_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} \tag{1.2}$$

or $\phi(B)e_t = \theta(B)a_t$ where $a_t$ are assumed to be independent Normal $(0,\sigma^2)$. Using the representation $\pi(B)e_t = a_t$, where $\pi(B) = \phi(B)/\theta(B)$, and assuming the parameters $\beta = (\phi,\theta)$ to be known, efficient estimates of the regression coefficients $\alpha$ can be found to a first approximation (i.e. neglecting end effects) by applying ordinary least squares (OLS) to $y' = x'\alpha + a$, where $y'_t = \pi(B)y_t$ and $x'_{k,t} = \pi(B)x_{k,t}$.

In theory, provided $e_t$ is stationary, application of OLS directly to $y$ and $x$ provides consistent estimates of $\alpha$. If further the variance of $x_{k,t}$ is concentrated in a relatively narrow frequency band, non-overlapping for each $k$, the actual estimates $\alpha_k$ may not be very sensitive to the operator or filter $\pi(B)$ that is chosen, whether or not it corresponds to the

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

true error structure. In other cases, the estimates of $\alpha$ may depend markedly on the choice of $\pi(B)$. In all cases, the standard error estimates for $\hat{\alpha}$ will be sensitive to the error model (as represented by $\pi(B)$) and also, it follows, will be any assessment of significance. The effect can be quite dramatic, for instance when the alternatives are $e_t = a_t$ and $\nabla e_t = a_t$, corresponding to OLS applied to $y$ and $x$ in the first case, and between $\nabla y$ and $\nabla x$ in the second - see Box and Newbold (1971). If correct inference about $\alpha$ is to be made, the error model for $e$ should therefore be chosen with care and its parameters $\alpha$ estimated efficiently - even though they may be regarded as nuisance parameters with respect to the regression.

We shall therefore tend to concentrate in this paper on estimation of the time series model for the error process. We shall be particularly concerned with investigating the presence of a random walk component in the error, because this can have a substantial influence on the estimated precision of the regression. In section 2 we indicate how the presence of regression terms in the model (1.1) can distort the error structure and lead to incorrect conclusions if least squares criteria are used without caution. Sections 3 and 4 introduce marginal likelihood as an estimation criterion which is capable of overcoming this distortion by treating the regression coefficients as nuisance parameters. Section 5 considers the particular cases of estimating first order autoregressive and first order moving average errors when the regression is simply a model constant term. Section 6 considers estimation of seasonality, section 7 tests of autoregressive roots of unity, section 8 intervention modelling, and section 9 the estimation of sinusoidal regression components, i.e. a mixed spectrum.

## 2. DISTORTION OF RESIDUALS BY REGRESSION.

A reasonable procedure for investigating the error structure is first to carry out OLS regression of $y$ on $x$, or of $\nabla y$ on $\nabla x$ if differencing is thought to be appropriate. The residual series is then taken as an estimate of $e_t$ or $\nabla e_t$, and subjected to various tests for autocorrelation, and possibly the full ARIMA model identification process as presented by Box and Jenkins (1976).

The possible distorting effect of the regression upon the autocorrelation has long been appreciated. Thus when the regression corresponds simply to mean correction of $y$, it is pointed out by Kendall and Stewart (1968) p. 435 that 'there is a downward bias in $r_j$, even for a random series', and when the true error structure is AR(1), that 'the bias in all these cases is downwards and obviously may be quite serious'. Correlograms which should decay (in the mean) towards zero in a geometric manner, tend to swing to negative values over a range of lags. The nature of the bias may be seen in the way the residual spectrum is affected. Mean correction removes power at frequency $\omega = 0$ which may have a large effect on AR(1) models with positive autocorrelation at lag 1, and on MA(1) models with negative autocorrelation, since much of the information about the parameters in these models occurs at low frequencies. Regression upon sinusoidal functions, and similarly upon indicator variables for seasonality, distorts the residual spectrum at other frequencies, and affects inference about the error structure as we see later.

Durbin and Watson (1950, 1951) focused attention upon the effects of regression when devising their test for autocorrelated error. The vast amount of subsequent work in this area is the subject of a valuable review by King (1983). He reports that likelihood ratio (LR) tests have relatively poor power against an AR(1) error structure with positive autocorrelation when the

regressors 'are smoothly evolving', and blames small sample bias of the AR parameter. He cites other authors who find the LR test unreliable.

Perhaps the simplest demonstration of how inference may be affected is to consider the case of level and trend estimation under the alternatives that the errors are random, or perform a random walk. Let us take

$$\text{Data } y_1 \cdots y_n$$

$$\text{Model 1} \qquad y_t = c + bt + e_t \; ; \quad e_t = a_t$$

$$\text{Model 2} \qquad y_t = c + bt + e_t \; ; \quad \nabla e_t = a_t$$

$$\text{or} \qquad \nabla y_t = b + a_t \; .$$

It will be useful later if we note here that under model 1 the OLS estimates come from solving

$$\begin{bmatrix} \Sigma y_t \\ \Sigma t y_t \end{bmatrix} = \begin{bmatrix} n & n(n+1)/2 \\ n(n+1)/2 & n(n+1)(2n+1)/6 \end{bmatrix} \begin{bmatrix} \hat{c}_1 \\ \hat{b}_1 \end{bmatrix} \; . \tag{2.3}$$

Under model 2 we need a constraint such as $e_1 = 0$ to estimate $\hat{c}_2 = y_1 - \hat{b}_2$, but OLS can be applied to $\nabla y_t$ for $\hat{b}_2$

$$\Sigma \nabla y_t = y_n - y_1 = (n-1)\hat{b}_2 \; . \tag{2.4}$$

The residual mean squares of these regressions may be calculated as $RMS_1 = (n-2)^{-1}\Sigma(y_t - \hat{c}_1 - \hat{b}_1 t)^2$, $RMS_2 = (n-2)^{-1}\Sigma(\nabla y_t - \hat{b}_2)^2$, the residual degrees of freedom being $(n-2)$ in both cases.

It would not be unnatural to select between these two models on the basis of the RMS values. It is however possible to evaluate their expected values under the two models; as shown in Table 1.

| True model | 1 | 2 |
|---|---|---|
| $E(RMS_1)$ | $\sigma^2$ | $(n+2)\sigma^2/15$ |
| $E(RMS_2)$ | $2n\sigma^2/(n-1)$ | $\sigma^2$ |

Table 1: Expected residual mean squares for models 1 and 2

Thus although for large samples, selection on the basis of the smallest RMS value would lead to acceptance of the correct model, we note that for $n < 13$ this criterion is biased in favour of model 1 when in fact model 2 is correct. In other words an ordinary linear regression fits a random walk better than a random walk model does! The bias is particularly galling in the case $n = 3$ when in fact $RMS_2 = 3RMS_1$, i.e. they are proportional statistics, and in fact there is no basis for discrimination.

The effect is no doubt worse when polynomials of higher degree are fitted, and may explain the attraction of prediction methods which fit, say, cubics to the latest segment of a series. Although for a random walk such fits will be 'good', the model is quite wrong and gives poor predictions.

The ratio $RMS_1/RMS_2$ does in fact have excellent properties for model discrimination provided that one does not simply examine whether it is greater or less than one, but instead takes the trouble to investigate its distribution. This is appreciated for example by Dicke (1978), who fits a trend line to the sequence of times of peak sunspot activity shown in Figure 1, and compares the observed RMS ratio which is .87, with the ratio of expected RMS values which is .46 under model 1 and 1.72 under model 2. In fact he uses the estimate $\hat{b}_1$ in both RMS quantities but this has relatively little effect on the formula. He takes the observed value of .87 as being more compatible with model 1. I am however grateful to Maxwell King for supplying me with the 1% critical value of .83 for the RMS ratio under model 1, and it would appear that the observations are compatible with neither model 1 nor model 2. The situation is complicated by the introduction of another regression variable, the peak sunspot number associated with each time point. We later reconsider this data.

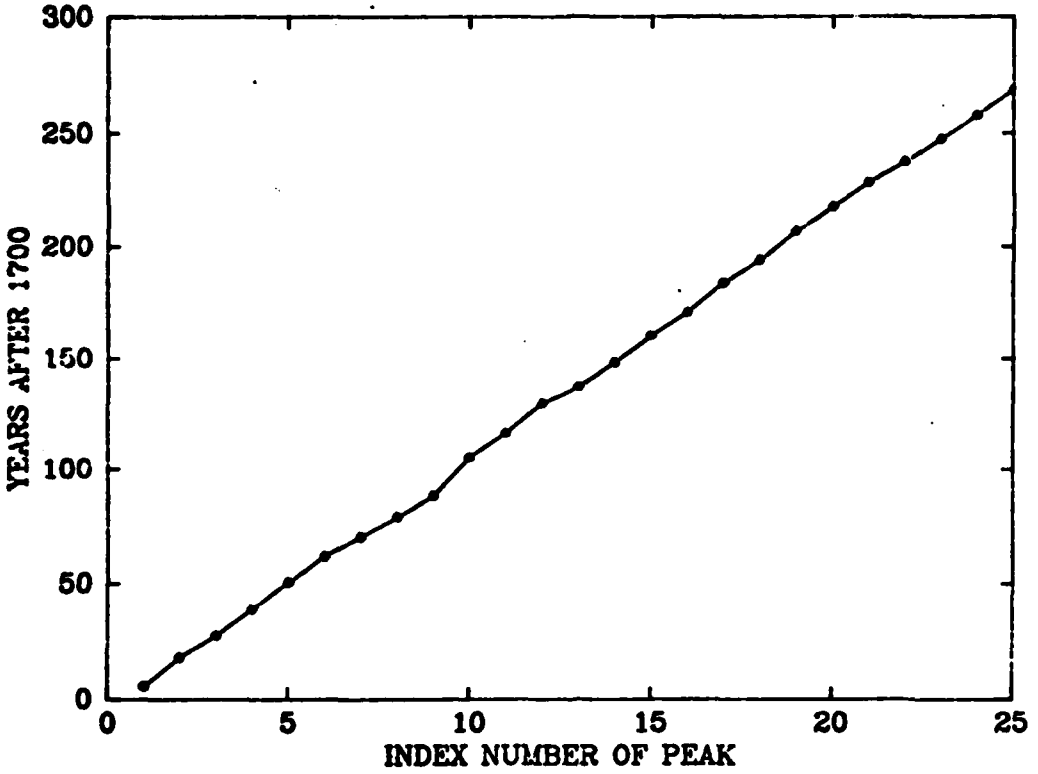



Figure 1: Peak years of Sunspot numbers.

The same problem, of testing for residuals which follow a random walk, is considered by Sargan and Bhargava (1983). Although exact tests against a specific autocorrelated error structure are valuable, the aim of this paper is to stress model estimation using a criterion which might be relied upon in general situations to indicate the true model.

## 3. ESTIMATION CRITERIA

We start with the distribution of the data for the model (1.1) and this is given formally by the density

$$f(y|\alpha,\beta,\sigma) \propto \sigma^{-n} M^{-1/2} \exp\{-1/2\ Q/\sigma^2\} \qquad (3.1)$$

where $Q = e'V^{-1}e$ is a quadratic form in the errors $e = y-x\alpha$, and the matrix $V$ has elements $V_{ij} = \text{cov}(e_i,e_j)/\sigma^2 = \gamma_{i-j}/\sigma^2$ which are functions of

$\beta$ only, $\gamma_k$ being the theoretical autocovariance sequence of the error model. Also $M = \det V$.

In practice $Q$ is calculated without explicitly constructing $V$ and is usually dominated by a sum of squares term $\Sigma a_t^2$ where $a_t$ is regenerated from $e_t$. E.g. for the ARMA(1,1) model, using the fact that $(a_1 - e_1)$ is independent of $a_1,\ldots,a_n$ we get

$$Q = \min_{a_1}\{\frac{(1-\phi^2)}{(\phi-\theta)^2}(a_1-e_1)^2 + \sum_1^n a_t^2\} \qquad (3.2)$$

where for $t = 2\ldots n$, $a_t = e_t - \phi e_{t-1} + \theta a_{t-1}$. There is some computational advantage in defining, somewhat artificially, $e_0 = a_0 = (a_1-e_1)/(\phi-\theta)$ which allows $t$ to run from 1 to $n$, and the minimization to be carried out wrt $a_0$.

Furthermore,

$$M = 1 + \frac{(\phi-\theta)^2}{(1-\phi^2)}(1+\theta^2+\cdots+\theta^{2n-2}) = 1 + \begin{cases} \frac{(\phi-\theta)^2}{(1-\phi^2)}\frac{(1-\theta^{2n})}{(1-\theta^2)} & \text{for } \theta = 1 \\ n(1-\phi)/(1+\phi) & \text{for } \theta = 1 \end{cases} \qquad (3.3)$$

This example will be useful in following illustrations.

Use of the exact likelihood function $L(\alpha,\beta,\sigma|y) = f(y|\alpha,\beta,\sigma)$ has now been supported by several estimation studies in the case of univariate models without regression components, e.g. by Ansley and Newbold (1980) who do not even introduce a constant term in the model. When regression components are present, inferences about $\alpha$ for any given $\beta$ can be carried out by direct maximization of $L$ using generalized least squares, since $e$ is linear in $\alpha$. This yields an estimate $\hat{\alpha}$ with dispersion matrix $\sigma^2 D$, where $D$ depends only upon $x$ and $\beta$ - we shall write $D(\beta)$. Thus given $\beta$ and $\sigma^2$, $(\hat{\alpha}-\alpha)$ is $MVN(0,\sigma^2 D(\beta))$.

Furthermore, maximization wrt $\sigma$ gives $\hat{\sigma}^2 = Q(\hat{\alpha},\beta)/n$, where $Q(\hat{\alpha},\beta)$ is the minimum of $Q$ w.r.t. $\alpha$. For later convenience we replace $n$ by $n-K$ in the definition of $\hat{\sigma}^2$, and this does not affect the result that the maximized likelihood is now

$$L(\beta) \propto M^{-1/2} Q^{-n/2} = \{M^{1/n} Q\}^{-n/2} \quad . \tag{3.4}$$

We shall call the quantity $M^{1/n}Q$ the deviance, $Dev(\beta)$. It may be regarded as an objective function to be minimized in the estimation of $\beta$ by maximum likelihood. In general, $\beta$ is a 'nonlinear' parameter so optimization methods are required. The strength of maximum likelihood rests very much on its proven reliability in many circumstances for comparisons within a parametric family of models. We now reconsider the problem from section 2 by viewing the two models as cases of a common parametric model:

$$\text{Model 3} \qquad y_t = c + bt + e_t \; : \; \nabla e_t = (1-\theta B)a_t$$

$$\text{or} \qquad \nabla y_t = b + (1-\theta B)a_t \quad .$$

The IMA(1,1) model for the error allows Model 1 in the case $\theta = 1$ and Model 2 in the case $\theta = 0$. Using formula (3.3) with $\phi = 0$ then gives the general likelihood from which we get (replacing $n$ by $n-1$)

$$\text{for Model 1} \quad (\theta=1) \quad Dev \propto n^{1/(n-1)} RMS\ 1$$

$$\text{for Model 2} \quad (\theta=0) \quad Dev \propto \quad RMS2 \quad .$$

We now note a penalty attached to Model 1 which was not present when we evaluated the likelihood, i.e. RMS1, without the differencing feature in the error structure. This penalty exists despite the fact that the differencing is 'cancelled' by the moving average operator when $\theta = 1$.

The introduction of this penalty was noted by Harvey (1980) when considering the problem of discriminating between regressions in levels and first differences. He used a general regression variable $x_t$ in place of the trend $t$, which we shall see is at the root of the problem in this case.

The penalty $n^{1/(n-1)}$ when applied to RMS1 helps to ameliorate the bias in the comparison, but does not overcome it until $n = 10$. Noting that Model 3 effectively eliminated the parameter $c$, we continue by imposing a further differencing structure on the error, to eliminate $b$;

$$\text{Model 4} \qquad y_t = c + bt + e_t \;\; ; \;\; \nabla^2 e_t = (1-\theta_1 B-\theta_2 B^2)a_t$$

$$\text{or} \quad \nabla^2 y_t = (1-\theta_1 B-\theta_2 B^2)a_t \;\; .$$

The IMA(2,2) model for the error now allows Model 1 in the case $\theta_1 = 2$, $\theta_2 = -1$, and Model 2 in the case $\theta_1 = 1$, $\theta_2 = 0$. It may now be shown that the exact likelihoods corresponding to these are given by the mean deviances

$$\text{for model 1} \qquad \text{Dev}/(n-2) = \{n^2(n^2-1)/12\}^{1/(n-2)}\,\text{RMS1}$$

$$\text{for model 2} \qquad \text{Dev}/(n-2) = (n-1)^{1/(n-2)}\,\text{RMS2} \;\; .$$

This gives a net penalty factor of $\{n^2(n+1)/12\}^{1/(n-2)}$ against Model 1. With these factors we can show that

$$E(\text{Dev 1}) = r\, E(\text{Dev 2}) \quad \text{when model 2 is true}$$

and

$$E(\text{Dev 2}) = s\, E(\text{Dev 1}) \quad \text{when model 1 is true}$$

where both $r$ and $s$ are greater than 1 for $n > 3$. When $n = 3$, Dev 1 $\equiv$ Dev 2. This criterion for model selection now seems satisfactory as Table 2 shows.

| n | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|
| r | 1.0328 | 1.083 | 1.142 | 1.205 | 1.271 | 1.329 | 1.407 |
| s | 1.0328 | 1.077 | 1.121 | 1.162 | 1.199 | 1.232 | 1.263 |

Table 2: Ratios of expected deviances under models 1 and 2.

Even for large $n$ the above net factor is important, implying that RMS1 should be inflated by 12% when $n = 100$ and 7% when $n = 200$.

## 4. MARGINAL LIKELIHOOD

The device of overdifferencing that seemed to solve the above problem evidently worked because it removed the question of regression parameter estimation. Very similar problems in the estimation of variance components models have been treated by using marginal likelihood, and Cooper and Thompson (1977) have applied these ideas to time series models. The fundamental idea is to evaluate the likelihood of the ARMA error structure parameters $\beta$, using only that information in $y$ which is invariant to any changes in the regression coefficients $\alpha$. This is what differencing achieved in our problem. In general the marginal likelihood can be evaluated in simple steps, for a proposed parameter value $\beta$.

(i) Estimate $\alpha$, obtain $Q(\hat{\alpha}|\beta)$ and $N = 1/\det D(\beta)$

(ii) The marginal deviance is $Mev(\beta) = \{NM\}^{1/(n-K)}Q$ where $M$ is the same factor as appeared in the exact likelihood. In practice $N$ is evaluated as the determinant of the matrix appearing in the GLS equations for $\hat{\alpha}$, so is obtained as a by product of solving these equations.

The expression given by Cooper and Thomson is equivalent to $Mev(\beta)$ apart from the fact that they effectively scale $N = N(\beta)$ by dividing by $N(0)$. The advantage is that $Mev$ is then invariant to a reparameterization of the regression by setting say $\tilde{\alpha} = P\alpha$, $\tilde{x} = xP^{-1}$. We shall omit this scaling, though bearing it in mind if we are tempted to compare models with different regressors. The advantage is that as defined in (ii), $Mev$ is invariant, apart from the usual Jacobian, to linear data transformations $(\tilde{y},\tilde{x},\tilde{e}) = R(y,x,e)$. Differencing and other simplifying operations such as are

considered by Abraham and Box (1978) are the most usual of such transformations.

Applying the above definition of Mev in the case of Models 1 and 2 leads very simply to the factors modifying RMS1 and RMS2 as the determinants of the least squares equation matrices in (2.3) and (2.4).

The expression for the Marginal Likelihood is obtained by the usual arguments. Because $\hat{\alpha}$ and $\hat{\sigma}^2$ are sufficient for $\alpha$ and $\sigma^2$ (and are independent), the marginal likelihood which represents information in $y$ invariant both to translations of $\alpha$ and scale change of $\sigma$, is given by $f(y|\hat{\alpha},\beta,\hat{\sigma})$.

Using the fact that $(\hat{\alpha}-\alpha)/\sigma$ is MVN[0,D(β)] and $(n-K)\hat{\sigma}^2/\sigma^2$ is chi-squared on (n-K)d.f., we can factor out the densities of $\hat{\alpha}$, $\hat{\beta}$ from (3.1) to obtain

$$f(y|\hat{\alpha},\beta,\hat{\sigma}) \propto (\text{Mev})^{-(n-K)/2} . \qquad (4.1)$$

We shall also use the fact that $(\hat{\alpha}-\alpha)/\hat{\sigma}$ has a multivariate T distribution with matrix parameter D(β), or MVT[D(β)].

The derivation by Cooper and Thompson is somewhat different, and the above corresponds more closely to that presented by Levenbach (1972) for auto-regressive models. The use of marginal likelihood in this kind of situation is in accord with the ideas of Kalbfleisch and Sprott (1970). Cooper and Thompson also mention that it is possible to derive (2.5) by Bayesian arguments. Taking the usual uniform priors on $\alpha$ and log $\sigma$ gives the same result after integrating out $\alpha$ and $\sigma$. It is then useful to interpret (4.1) as $f(\beta|y)$, the posterior density for $\beta$ given a uniform prior. The same distributional property as above still holds for $(\hat{\alpha}-\alpha)/\hat{\sigma}$, but now with the interpretation that it specifies $f(\alpha|\beta,y)$. Zellner and Tiao (1964) present the Bayesian approach for the case of AR(1) error structure.

Because $Mev(\beta) = \{NM\}^{1/(n-K)} \min_{\alpha} Q(\alpha,\beta)$, and neither N nor M depend on $\alpha$, it is tempting to take $NM^{1/(n-K)}Q(\alpha,\beta)$ as an overall objective function to be minimized wrt $\alpha$ and $\beta$. Although this may be useful from a computational viewpoint, the interpretation of this function is difficult and its use in tests which simultaneously involve $\alpha$ and $\beta$ is not recommended. We shall however be so bold as to use contours of $Mev(\beta)$ to define confidence regions and carry out mean deviance tests by referring $\{Mev(\beta)-Mev(\hat{\beta})\}/\{Mev(\hat{\beta})/(n-K-L)\}$ to chi-squared on L d.f. where $\hat{\beta}$ is the minimum deviance estimate of $\beta$, and L is the number of ARIMA model parameters. If differencing is present in the ARIMA model, K includes the reduction in error series length due to the differencing.

## 5. CONSTANT TERM ESTIMATION

Considering the error models $e_t = a_t$ (random error) and $\nabla e_t = a_t$ (random walk error) as two extremes, it is possible to bridge these extremes by either of two simple models, the AR(1) model $e_t = \phi e_{t-1} + a_t$, and the IMA(1,1) model $\nabla e_t = a_t - \theta a_{t-1}$. When $\phi = 0$ or $\theta = 1$ we obtain the random error, when $\phi = 1$ or $\theta = 0$ we obtain the random walk.

Introducing the constant term in the AR(1) case corresponds to a constant term in the regression. Introducing a constant in the IMA(1,1) case is equivalent to a trend regression for y, i.e. $y_t = ct + e_t$, whence $\nabla y_t = c + a_t - \theta a_{t-1}$. Such models occur frequently in practice.

The marginal likelihood for AR(p) models with a constant term was considered by Levenbach (1972). He gave some explicit formulae and contour maps for the marginal likelihood. In the case of the AR(1) model the contrast is between $Dev = (1-\phi^2)^{-1/n}Q$ for the exact likelihood, and $Mev = \{1+n(1-\phi)/(1+\phi)\}^{-1/(n-1)}Q$ for marginal likelihood. The factor in the former

is well known, and corresponds to (2.3) when $\theta = 0$. The factor in the latter may be got by the device of taking the exact likelihood of the differenced series, and thus corresponds to (2.3) when $\theta = 1$. The main point to note is that the exclusion of $\phi = 1$ in the exact likelihood is absent in the marginal likelihood. A valid comparison of the random walk model corresponding to $\phi = 1$, can then be made with the AR(1) model with $\phi < 1$. In the limit as $\phi \to 1$, Mev $\to \sum_{2}^{n} (\nabla y_t)^2$, but there is numerical instability in the calculation of Mev from the AR(1) model as $\phi \to 1$. Both in the calculation of $\hat{c}$ which has a well defined limit of $(y_1 + y_n)/2$, and of the product NM, ratios of quantities with the common factor $(1-\phi)$ occur. It is possible for Mev to be decreasing with strictly negative gradient up to $\phi = 1$. In fact at $\phi = 1$, $\partial \log \text{Mev}/\partial\phi = 1/2\{1 - (y_n - y_1)^2 / \sum_{2}^{n} (\nabla y_t)^2\}$. Although this has expected value zero if $y_t$ is indeed a random walk, it may be negative if $y_t$ has a strong trend.

The marginal likelihood for the MA(1) model with constant term was the subject of simulation studies by Cooper and Thomson (1977). The quantity Dev in this case has as the multiplier of Q, the factor $M^{1/n}$, where using (2.3) with $\phi = 0$, we have:

$$M = (1-\theta^{2n+2})/(1-\theta^2) \quad \text{for } \theta^2 \neq 1, \quad \text{and } M = n \quad \text{for } \theta^2 = 1 .$$

The multiplier of Q in Mev is $(MN)^{1/(n-1)}$ where

$$MN = \begin{cases} [n(1-\theta^{2n+2})/(1-\theta^2) - (1-\theta^{n+1})^2/(1-\theta)^2]/(1-\theta)^2, & \text{for } \theta^2 < 1 \\ n^2(n^2-1)/12 , & \text{for } \theta = 1 \\ n^2/4 \text{ when } n \text{ is odd, } (n^2-1)/4 \text{ when } n \text{ is even,} & \text{for } \theta = -1 . \end{cases}$$

Symmetry considerations show that both Dev and Mev always have stationary points at the boundaries, $\theta = \pm 1$, so that minimum Dev or Mev values can occur on the boundaries.

Note that, as for the AR(1) model, the value of MN is n at the origin i.e. $\theta = 0$ in this case, so the effect on minimum Mev estimation is seen by comparison with this value. The value $\theta = 1$ is much more strongly discouraged by MN than by M alone, whereas $\theta = -1$ is slightly less discouraged.

For both the AR(1) and MA(1) models, the use of marginal likelihood counteracts the tendency of mean correction to bias sample autocorrelation downwards. For the MA(1) model with $0 < \theta < 1$ it is known, see Cryer and Ledolter (1981), that the estimates $\hat{\theta}$ have a tendency to pile up at $\hat{\theta} = 1$ even when the mean level is not estimated, but set to its true value. Thus Cooper and Thompson report that out of 200 simulations of a series of length $n = 59$ with $\theta = .7539$, the value $\hat{\theta} = 1$ occurred 8 times. When the series mean was estimated, still using exact likelihood, this leapt to 65 times out of the 200. When marginal likelihood was used it fell back to 19 times.

The distinction between $\theta = 1$ and $\theta < 1$ is important in the model $y_t = y_{t-1} + c + a_t - \theta a_{t-1}$ which has a structural interpretation for $y_t$ as a signal plus noise, the signal being a random walk. Thus $y_t = w_t + n_t$ where $w_t = w_{t-1} + c + \alpha_t$ and $n_t = \beta_t$, both $\alpha_t$ and $\beta_t$ being independent Normal with zero means and variances $\sigma_\alpha^2$, $\sigma_\beta^2$. The variance ratio $r = \sigma_\alpha^2/\sigma_\beta^2$ is related to $\theta$ by $(1-\theta)^2/\theta = r$, the value $\theta = 1$ corresponding to absense of the random walk component, apart from a fixed increase c per unit time. If $\theta = .7539$ then $r \sim 1/12$ and over a series length 60 the random walk component will build up a variance of 5 times the noise

level, and should be detectable. The simulations of Cooper and Thompson show how frequently the wrong conclusion that $\theta = 1$ can arise in this situation when $c$ is estimated, unless marginal likelihood is used.

We now move on to consider particular data sets which illustrate how the use of marginal likelihood can affect inference.

## 6. ESTIMATION OF SEASONALITY

We consider the selection between a model which represents seasonality using fixed seasonal coefficients applied to indicator variables, and one in which the Box-Jenkins seasonal model is used. A fixed linear trend is also estimated in both models. We shall use the series of 144 points of the logarithms of airline passenger totals as analyzed by Box and Jenkins (1976) and shown in Figure 2.

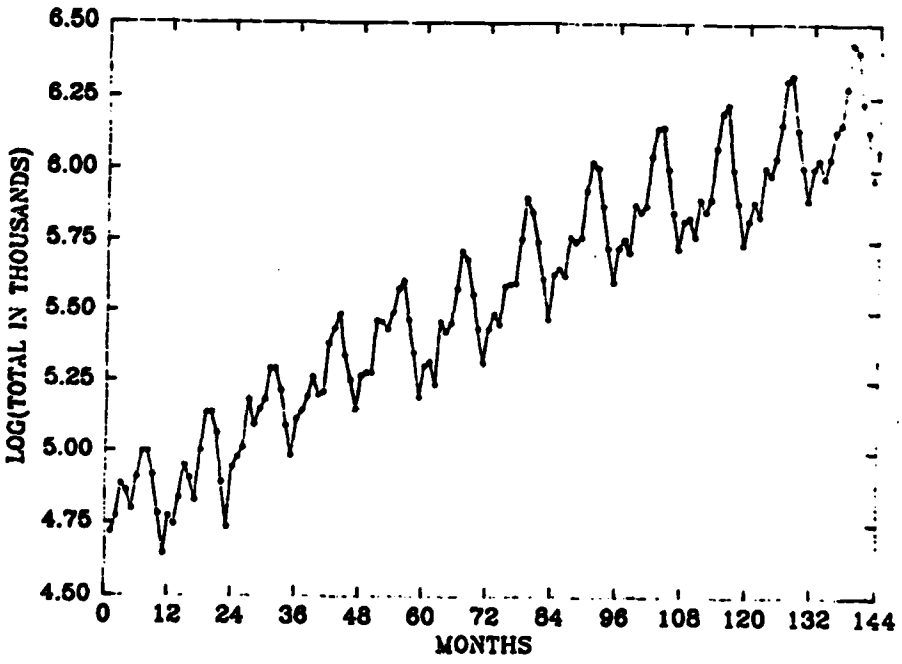



Figure 2: Airline Passenger Totals

Let us take:

model 1: $y_t = bt + \sum_{j=1}^{12} M_j S_{j,t} + e_t$

where for $j = 1...12$, $S_{j,t} = 1$ in month $j$, else is 0.

model 2: $y_t = bt + \frac{(1-\Theta B^{12})}{(1-B^{12})} e_t$ ,

so that in fact model (ii) with $\Theta = 1$ is in theory identical to (i).

The error $e_t$ from an OLS fit of model 1 is evidently strongly autocorrelated and as a first step we represent it by an AR(1) model, $(1-\phi B)e_t = a_t$. Then model (i) is a fairly classical decomposition model.

The results of fitting these models using exact likelihood are summarized in rows 1 and 2 of Table 3. We note that the minimum Dev occurs with the fixed seasonality model 1. Reports to the effect that such fixed seasonality models fit better than Box-Jenkins models have been heard by the author for some years. The fitting criterion has usually been least squares, i.e. minimization of $Q$ rather than $Dev = M^{1/n}Q$, but this is very similar to exact likelihood at parameter values such as the above, which are reasonably well distant from the boundaries. Nevertheless some puzzlement has resulted, because on the L.S. criterion model 2 is more general than model 1, and should obtain a residual sum of squares at least as good as model 1. The explanation is that the least squares surface for model 2 has, besides a local minimum at $\hat{\Theta}$, a sharp dip at $\Theta = 1$ to what in this case is a lower point corresponding to model 1. The exact likelihood multiplier $M$ fills this dip, and rather fortuitously the back-forecasting method of Box and Jenkins if insufficiently iterated (as in the early software) has a similar effect which avoids $\Theta = 1$. Marginal likelihoods however give proof to support the use of Box-Jenkins seasonal models which may formerly have required some faith.

| Model | Dev or Mev | RMS | RDF | $\hat{b}$ | SE($\hat{b}$) | $\hat{\phi}$ | $\hat{\theta}$ | $\hat{\Theta}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | .17721 | .001354 | 130 | .00999 | .00032 | .787 | | |
| 2 | .18874 | .001401 | 129 | .00996 | .00058 | .779 | | .577 |
| 3 | .18421 | .001416 | 130 | .00950 | .00238 | | .262 | |
| 4 | .18296 | .001369 | 129 | | | | .402 | .557 |
| 1(M) | .23901 | .001354 | 130 | .00999 | .00035 | .803 | | |
| 2(M) | .20110 | .001403 | 129 | .00995 | .00063 | .792 | | .568 |
| 3(M) | .24009 | .001416 | 130 | .00949 | .00233 | | .261 | |

Table 3: Summary of models fitted to the 'airline data'

Fitting Models 1 and 2 by marginal likelihood criteria gives the results shown in rows 1(M) and 2(M) of Table 3. The picture has changed considerably and model 2(M) is now much more strongly favoured than 1(M). Although the parameter values are little changed, a heavier 'penalty' has been introduced to model 1(M), arising from the use of seasonality indicators.

We continue by considering the model which was actually used by Box and Jenkins for this data. It involved a further stage of differencing by taking for the error structure

$$e_t = \frac{(1-\theta B)}{(1-B)} a_t \quad .$$

Using this in models 1 and 2 gives

$$\text{model 3:} \quad \nabla y_t = b + \Sigma M_j S_{j,t} + (1-\theta B)a_t \quad ;$$

$$\text{model 4:} \quad \nabla\nabla_{12} y_t = (1-\theta B)(1-\Theta B^{12})a_t \quad ;$$

the last one being of course the Box-Jenkins model.

Estimation by exact likelihood gave the results in rows 3 and 4 of Table 3. Although the Box-Jenkins model 4 is now slightly preferred to models 2 and 3, it still takes second place to model 1. Again however, looking at the results for marginal likelihood, we see that model 4 - for which exact and marginal likelihood coincide - is now much to be preferred. As a check, model 4 was refitted with $\theta$ constrained to 1.0, which should be equivalent to model 3(M) and did in fact give the same value of Mev and estimate of $\theta$. Note incidentally how little the trend estimate $\hat{b}$ in Table 1 changes through models 1, 2 and 3, though its standard error estimate changes substantially.

## 7. DISCRIMINATING AR ROOTS OF UNITY

To illustrate the problem we first look at a data set of 200 points of the series of monthly unemployment in Scotland from January 1952 to August 1968, as shown in Figure 3. Two competing models are, after applying a fourth-root transformation to the data,

$$\text{model 1} \quad (1-\phi_1 B-\phi_2 B^2)\{\nabla_{12} y_t - c\} = (1-\theta B)(1-\Theta B^{12})a_t$$

$$\text{model 2} \quad (1-\phi_1 B)\nabla\nabla_{12} y_t = (1-\theta B)(1-\Theta B^{12})a_t \quad .$$

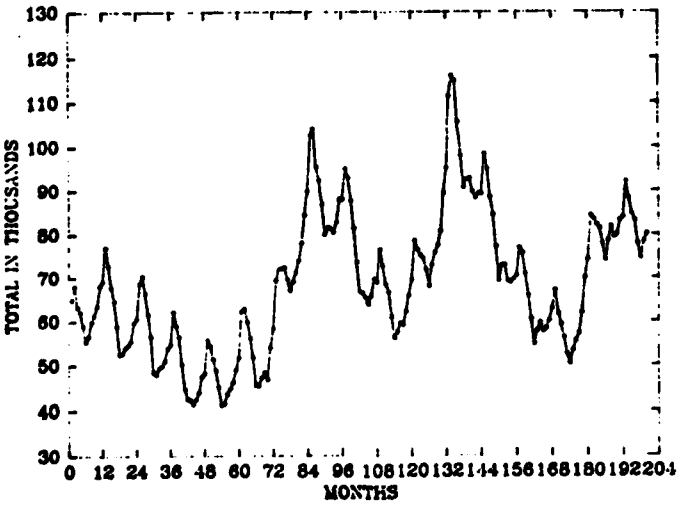



Figure 3: Unemployment in Scotland.

In model 1, as a root of $(1-\phi_1 B-\phi_2 B^2)$ approaches 1, i.e. $\phi_1 + \phi_2 \to 1$, the exact likelihood becomes 0, or the Deviance infinite. In this limit, model 1 becomes model 2, but although an exact likelihood for model 2 can be defined on the basis that $\nabla\nabla_{12}y_t$ is stationary, this is evidently not directly comparable with that for model 1. However, because of the presence of the constant c in model 1, the marginal likelihood may be used in this comparison, since it is based on the same information, i.e. that in $\nabla\nabla_{12}y_t$. The results from models 1(M) and 2 in Table 4 may be used in a 'mean deviance' test by referring

$$(1.107046 - 1.067656) \times 183/(1.067656) = 6.752$$

to the chi-squared distribution on 1 d.f., from which it might reasonably be inferred that model 1 is to be preferred, despite the fact that $\hat{\phi}_1 + \hat{\phi}_2 = .988 \simeq 1$. This is to a large extent confirmed by superior forecasts produced by model 1 over the next few years. The operator $(1-\hat{\phi}_1 B-\hat{\phi}_2 B^2)$ corresponds to a spectrum which peaks around a period of 5 years and results, in the long term, in a forecast function with a damped cyclical pattern. Model 1 is similar in the short term, but does not imply a turning point in the trend of its forecasts.

| Model | Dev/Mev | RMS | RDF | $\hat{\phi}_1$ | $\hat{\phi}_2$ | $\hat{\theta}$ | $\hat{\Theta}$ | $\hat{c}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 1.061129 | .005339 | 183 | 1.9128 | -.9250 | .6804 | .7727 | .0757 |
| 1(M) | 1.067656 | .005351 | 183 | 1.9086 | -.9201 | .6656 | .7664 | .0766 |
| 2 | 1.107046 | .005644 | 184 | .8709 | | .5381 | .7909 | |

Table 4: Summary of models fitted to unemployment data

It may be objected that the exact likelihood in row 1 of Table 4 is only slightly less than the marginal value in row 1(M), and would lead to the same conclusion. To some extent this is coincidence - even at the origin of the ARMA parameters, the definition of Mev is some 3% greater than that of Dev. A further example in this section shows that the exact likelihood for the 'undifferenced' model can exceed that for the differenced model, so that a 'mean deviance' quantity is negative and evidently meaningless.

In this next case then, we examine 150 points of annual measurements of day length variations as given by Luo, et al (1977) and shown in Figure 4. The data has been smoothed, which introduces MA terms in the model, but also has very much the appearance of a Random Walk (R.W.) and the model choice reflects this possibility

$$\text{model 1} \quad (1-\phi_1 B-\phi_2 B^2)(y_t-c) = (1-\theta_1 B-\theta_2 B^2)a_t$$

$$\text{model 2} \quad (1-\phi_1 B)\nabla y_t = (1-\theta_1 B-\theta_2 B^2)a_t \quad .$$

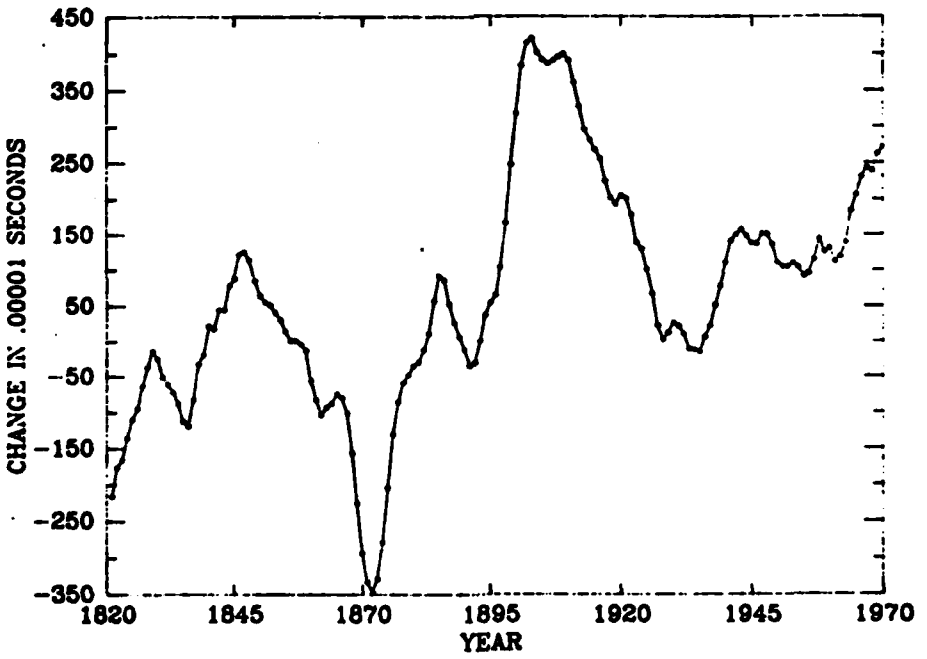



Figure 4: Changes in daylength

| Model | Dev/Mev | RMS | RDF | $\hat{\phi}_1$ | $\hat{\phi}_2$ | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{c}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 37990.88 | 250.48 | 145 | 1.3881 | -.4158 | -.5366 | -.6148 | 50.52 |
| 1(M) | 37044.50 | 250.55 | 145 | 1.3872 | -.4034 | -.5440 | -.6173 | 44.10 |
| 2 | 37154.62 | 251.44 | 146 | .3869 | | -.5542 | -.6207 | |

**Table 5: Summary of models 1 and 2 fitted to the day length data**

We first note that Dev for model 1 exceeds that for model 2 (for which DEV = MEV). The mean deviance test statistic from the results in rows 1(M) and 2 is now 0.431 and we might conclude that there is insufficient evidence to reject the RW type model 2. We do however take a further look at this data in section 9.

We now re-examine the peak-sunspot data of Figure 1, taking the model

$$y_t = c + bt + dx_t + e_t \quad ; \quad t = 1...25 \qquad (7.1)$$

where $x_t$ is the corresponding sequence of sunspot numbers at the peaks, and we take an AR(1) error model $e_t = \phi e_{t-1} + a_t$. Interest lies mostly in the estimate of b and its precision. When $\phi = 0$, i.e. with random error, we have $\hat{b} = 11.136$ (SE.042) and RMS 2.094 whereas with $\phi = 1$, i.e. random walk error, we have 11.051 (SE.302) and RMS 2.181. Estimating $\phi$ by exact likelihood gives $\hat{\phi} = .472$ (SE .216) and $\hat{b} = 11.113$ (SE .065), but using marginal likelihood $\hat{\phi} = .625$ (SE .217) and $\hat{b} = 11.102$ (SE .085). In such a situation the use of a point estimate for $\phi$ does not give the full picture, and in Figure 5 we show the marginal likelihood function for $\phi$ together with the estimates of b and of its standard error. Although $\phi = 0$ seems definitely ruled out, $\phi = 1$ still appears to be a possibility. Taking a Bayesian interpretation of the marginal likelihood as the marginal density of $\phi$, i.e. $f(\phi|y)$, and taking the $t_{22}$ density of $(\hat{b}-b)/SE(\hat{b})$ as specifying

the density $f(b|\phi,y)$ we can numerically construct the posterior for $b$ as $\int f(b|\phi,y)f(\phi|y)d\phi$. This is shown in Figure 6. The mode is at $b = 11.115$, a 95% confidence interval for $b$ is (10.815, 11.325) and a 99% interval is (10.62, 11.49). These are much wider than those found by taking $\phi$ fixed at .625. Further studies reveal that points 7, 8 and 9 of the data have a considerable influence and other error models which 'follow' these closely, e.g. MA(2), give a somewhat different answer. The answer to Dicke's question may therefore not be clear, but it usefully illustrates the statistical considerations.

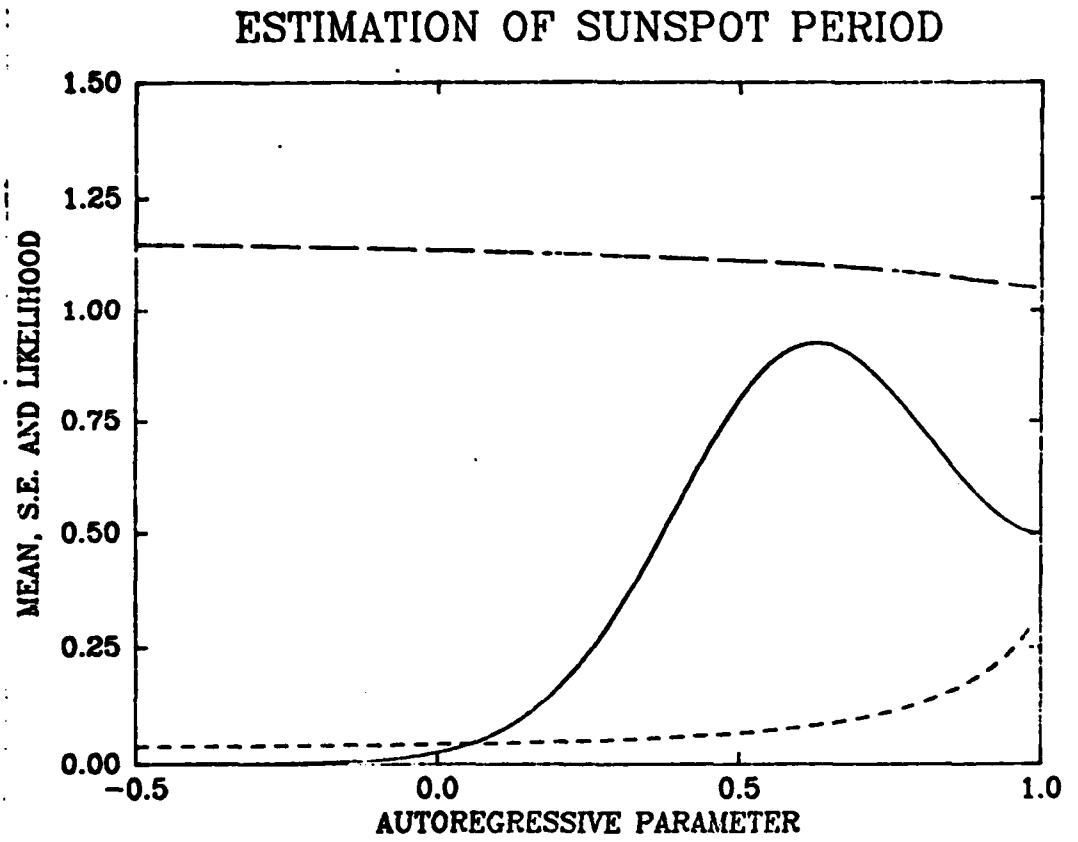



Figure 5: Marginal likelihood (solid line), estimated period of sunspot peaks minus ten years (long dashed line) and standard error of the period (short dashed line), as functions of the autoregressive parameter.

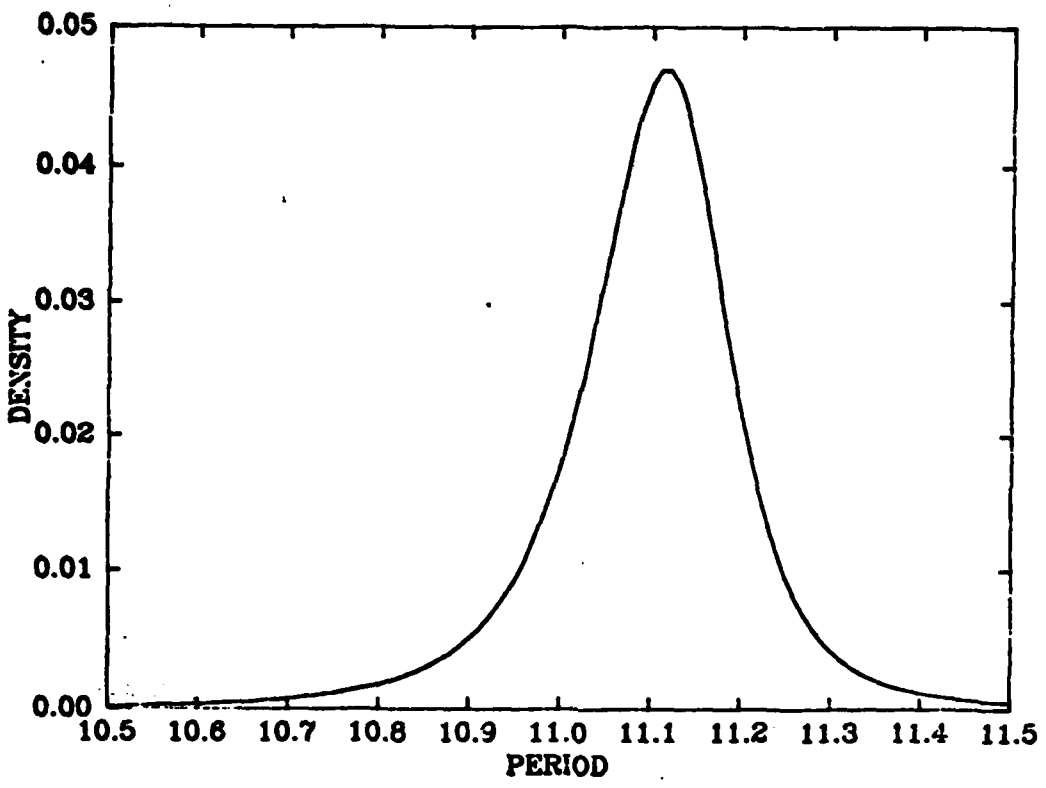



Figure 6: Marginal posterior density of the sunspot period

## 8. INTERVENTION MODELLING

A strong argument for the use of marginal likelihood may be found in the treatment of missing observations from time series data, and the effect on ARIMA parameter estimates. This can be cast as a regression problem by using intervention analysis techniques - see Box and Tiao (1975) - and we extend our consideration to such simple intervention patterns as step changes in the level of a series.

To appreciate the possible effect consider a series for which a simple MA(1) model is appropriate, and is fitted using the exact likelihood criterion:

$$\text{Dev} = \{(1-\theta^{2n+2})/(1-\theta^2)\}^{1/n} Q(\theta) \qquad (8.1)$$

where $Q(\theta) = \min_{a_0} S(\theta|y)$ and the sum of squares $S$ is given by

$$S(\theta|y) = \sum_{t=0}^{n} a_t^2 \quad \text{where} \quad a_t = y_t + \theta a_{t-1} \quad \text{for} \quad t = 1...n \ .$$

Suppose now $y_{r+1}$ is missing for some $0 < r < n-1$. It may be treated by introducing a regression (intervention) variable $I_t$ which is an impulse at $t = r+1$, after first setting $y_{r+1} = 0$. Then $\hat{y}_{r+1} = -\hat{\alpha}$ where $\hat{\alpha}$ is the estimated regression coefficient.

It is easily appreciated that in this case we can instead take $y_{r+1}$, or equivalently $a_{r+1}$ as the nuisance parameter so that $Q(\theta) = \min_{a_0, a_{r+1}} S(\theta|y)$ in the definition of Dev.

Now note that if $y_{r+1}$ is missing, we have for this model two independent samples $y_1,\ldots,y_r$ and $y_{r+2},\ldots,y_n$, of lengths $r$ and $s = n-r-1$, so that the true likelihood of the actual observations (forgetting the missing value) is found as the product of the likelihoods for the two parts. This leads to the following expression for the deviance which must of course equal the marginal deviance from the regression formulation, since it is based on just the available observations.

$$\text{Mev} = \{(1-\theta^{2r+2})(1-\theta^{2s+2})/(1-\theta^2)^2\}^{1/(n-1)} Q(\theta) \qquad (8.2)$$

where $Q(\theta)$ is exactly as defined above.

Incidentally, the interpolated value is $\hat{y}_{r+1} = \hat{a}_{r+1} - \theta\hat{a}_r = \hat{y}_F + \hat{y}_p$ where $\hat{y}_p$ is the forecast of $y_{r+1}$ based upon the first part of the data, or the past wrt $y_{r+1}$, and $\hat{y}_F$ is the back forecast based on the future part of the data. This is of course a direct consequence of the independence of the two parts for this model.

Assuming that Mev is calculated from the intervention regression formulation, the factor $N$ in the general expression for Mev, converts Dev in (8.1) into Mev in (8.2) for this example. Note that e.g. for a series of length $n = 100$, a missing value at $t = 60$ increases the

multiplier of Q from 1.047 for Dev to 1.082 for Mev at $\theta = 1$.

A more commonly occurring situation which is precisely equivalent, is when a step function is introduced to model a change in level at time r+1 in a series which follows an IMA(1,1) model, i.e.

$$y_t = S_t + \frac{(1-\theta B)}{\nabla} a_t$$

where $S_t = 0$ for $t \leq r$ and 1 for $t > r$. On differencing, we get $\nabla y_t = I_t + (1-\theta B)a_t$, the model just considered.

We now have the interpretation that in fitting such a step, one is removing some evidence for a random walk (signal) component in $y_t$, as discussed in section 5. The use of marginal likelihood acts against the corresponding tendency for $\theta$ to shift towards 1.

As an illustration, consider the slightly more complex case of 122 values from the latter part of the Beverage wheat price index as shown in Figure 7. An IMA(1,3) model appears sensible for this data after a logarithmic transformation, and a constant term is included because of the firm positive trend over the historical record of this series. At points $t = 70$ and 71 in our subseries, there appears to be a general shift down in the series level, following a slightly high value at $t = 69$. Three interventions are introduced; an impulse at $t = 69$ and steps at each of $t = 70, 71$. This has the effect of rendering the differenced section from $t = 2...68$ independent of that from $t = 72...122$, given the model. From Table 6 it is seen that without the intervention estimation the steps down are taken as evidence for the RW signal component, and there is no hint that $\theta(B) = 1-\theta_1 B-\theta_2 B^2-\theta_3 B^3$ has a root of $B = 1$. Estimations by Dev in row 1 or Mev in row 1(M) are then similar. However, estimation of the three intervention effects by exact likelihood as shown in row 2, leads to a root of $\theta(B)$ at one, implying absence of any random walk component. The use of marginal likelihood counters

this effect as seen in row 2(M). Although standard errors are not given, the t values were 1.82 for $\hat{I69}$, 1.09 for $\hat{S70}$ and 2.78 for $\hat{S71}$. So the only real effect seems to be associated with S71, although the drop from $y_{69}$ to $y_{70}$ appeared large.

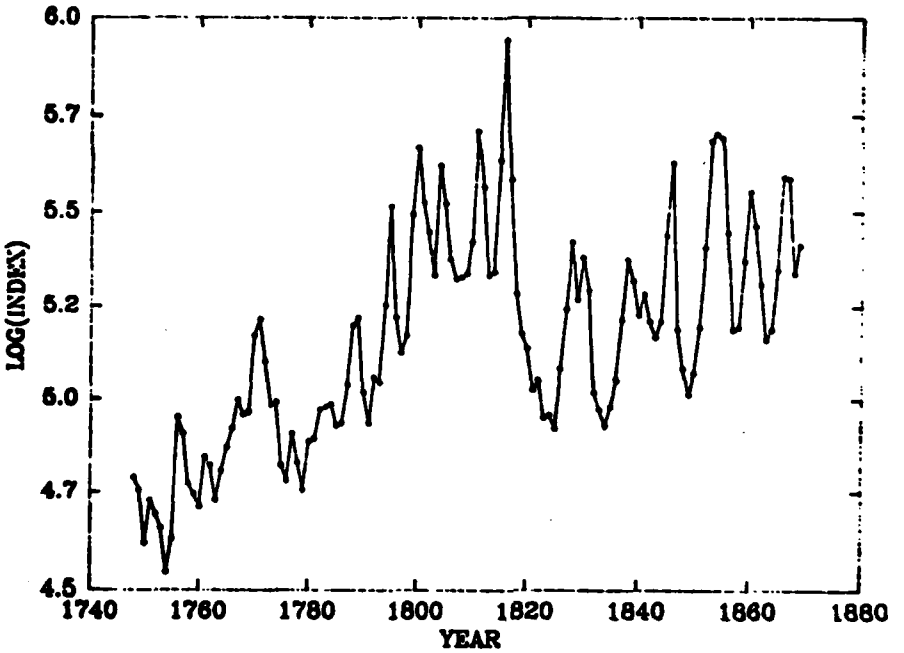



Figure 7: Beveridge Wheat Price Index

| Model | Dev/Mev | RMS | RDF | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}_3$ | $1-\Sigma\theta_i$ | $\hat{c}$ | $\hat{I69}$ | $\hat{S70}$ | $\hat{S71}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2.0971 | .01782 | 117 | -.059 | .406 | .314 | .339 | .006 | | | |
| 1(M) | 2.2189 | .01784 | 117 | -.073 | .390 | .303 | .380 | .006 | | | |
| 2 | 1.7284 | .01461 | 114 | .091 | .516 | .393 | .000 | .011 | .225 | -.155 | -.346 |
| 2(M) | 1.8864 | .01513 | 114 | .038 | .459 | .341 | .162 | .011 | .217 | -.160 | -.328 |

Table 6: Summary of models fitted to the wheat price index.

It must be remembered that these conclusions are affected by the choice of sample size. For short samples it will be more difficult to detect the RW component even using Mev, and it will be easy, even using Dev, in large samples. It can be expected however that over a good range of sample sizes, Mev may be found useful.

## 9. MIXED SPECTRUM ESTIMATION

The search for deterministic sinusoidal components in time series is still of great interest in many fields. The significance of such a discrete component, at a given frequency, depends upon how prominent the sample spectrum is at that point in comparison with the general level in the surrounding frequency band, which is supposed to arise from a continuous spectrum. If interest lies in a single component at a well defined frequency, well away from the ends of the frequency range, then the task is quite reasonable since the continuous part of the spectrum should be adequately represented and well estimated using an ARMA model - see Campbell and Walker (1977). If however there are several discrete components covering a frequency band, their estimation may deplete the sample spectrum over this band and distort estimation of the model of the continuous spectrum. This possibility is accentuated if the end of the frequency range is affected, because the ARMA model is required to extrapolate rather than to interpolate the missing part of the spectrum, which must be estimated if any reasonable inference is to be made. If a very flexible ARMA model is chosen (e.g. a large number of parameters), then estimation using criteria such as exact likelihood can be expected to make the model fit the low spectrum in the depleted band, leading to unrealistically high significance levels for the discrete components. What is required, is an ARMA model of limited flexibility, yet sufficiently

plausable as a representation of any true underlying continuous spectrum. This model should then be fitted only to the unaffected part of the spectrum, and its inter/extrapolation over the depleted range used to assess significance. This is precisely what Marginal likelihood achieves - it uses just that information which cannot be affected by estimation of the discrete components. The main danger is that there is insufficient information remaining to adequately determine some aspects (parameters or parameter combinations) of the ARMA model.

We again turn to the 'daylength' data in which discrete (sinusoidal) components have been found - see Luo et al, by searching for the highest peaks in the sample spectrum. One ought, because of this, to be much more severe in conceding significance of any coefficient, but as we shall see later this may not be a major factor in the inference problem in this case. We shall therefore treat the frequencies as fixed, and the corresponding periods are given in Table 7 with those of the fundamental and first seven harmonics of the data for comparison. In fact a trend plus twelve frequencies were used by Luo et al, but eight frequencies are quite sufficient to illustrate the points here. Although it is tempting, as a result of our previous model fitting to this data in section 6, to work with the differenced data, this would completely prejudice the outcome of the analysis by diminishing the large variance at low frequencies which is the focus of interest. We therefore retain the ARMA(2,2) model as the model for the error structure in a regression upon discrete components $\cos \omega_j t$ and $\sin \omega_j t$ for $t = 1 \ldots n=150$. Note that if the AR operator in the model has a root close to unity then we would expect much lower levels of significance for the fitted components. A trend term was also included to match the analysis of Luo et al. A sequence of estimations was then carried out, pairs of cosine and sine

components with coefficients $c_i$, $s_i$ being introduced for $i = 1,2,3...$, corresponding to the modeling periods in Table 7. We select the results of estimations with 4 such pairs, and with 8 such pairs, repeated for exact and marginal likelihood estimation. In the case of marginal likelihood estimation with 8 pairs, there were found to be two distinct local minima in the ARMA model parameter space, and the results for all these estimation points are summarized in Table 8. A "t" value is given besides each regression coefficient estimate, being the ratio to its estimated standard error.

| Component # | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Modeling period | 178.698 | 89.348 | 59.555 | 45.0 | 34.503 | 29.783 | 22.337 | 19.885 |
| Harmonic period | 150 | 75 | 50 | 37.5 | 30 | 25 | 21.43 | 18.75 |

Table 7: Periods of discrete sinusoidal components fitted to daylength series

| Model | 4 component (Dev) | 4 component (Mev) | 8 component (Dev) | 8 component (Mev) | 8 component (Mev)B |
|---|---|---|---|---|---|
| Dev/Mev | 30747.36 | 29027.36 | 21217.48 | 24372.40 | 24154.68 |
| RMS | 219.206 | 249.293 | 162.189 | 246.050 | 183.607 |
| RDF | 136 | 136 | 128 | 128 | 128 |
| $\hat{\phi}_1$(SE) | 1.239(.107) | 1.374(.111) | .990(.110) | 1.397(.122) | 1.077(.120) |
| $\hat{\phi}_2$(SE) | -.416(.105) | -.383(.117) | -.499(.101) | -.397(.142) | -.314(.126) |
| $\hat{\theta}_1$(SE) | -.508(.094) | -.557(.091) | -.437(.099) | -.548(.095) | -.557(.102) |
| $\hat{\theta}_2$(SE) | -.615(.082) | -.625(.079) | -.612(.085) | -.625(.079) | -.628(.081) |
| mean(t) | -690.4(2.44) | -778.2(1.54) | -1617.0(4.53) | -1195.5(0.59) | -1351.5(2.56) |
| trend(t) | 9.88(2.65) | 10.99(1.86) | 21.75(4.60) | 16.30(1.48) | 18.30(2.62) |
| c1(t) | 156.4(1.57) | 181.00(0.87) | 408.81(3.33) | 284.49(0.90) | 328.86(1.80) |
| s1(t) | 384.6(2.11) | 440.87(1.45) | 988.54(4.31) | 716.12(1.35) | 816.98(2.40) |
| c2(t) | 223.0(2.82) | 253.52(1.82) | 443.65(4.59) | 342.09(1.47) | 378.99(2.61) |
| s2(t) | 123.7(2.15) | 142.50(1.24) | 339.91(5.07) | 245.37(1.42) | 279.40(2.76) |
| c3(t) | -15.7(0.29) | -.853(0.01) | 138.93(2.17) | 67.41(0.42) | 94.29(0.96) |
| s3(t) | 55.4(2.04) | 60.71(0.97) | 122.39(7.12) | 100.42(1.39) | 109.97(3.62) |
| c4(t) | 52.8(1.64) | 61.95(1.04) | 145.56(4.65) | 106.45(1.20) | 122.11(2.42) |
| s4(t) | -53.8(2.04) | 54.73(1.06) | -54.94(2.93) | -47.37(0.74) | -49.43(1.53) |
| c5(t) | | | 58.35(3.06) | 35.56(0.59) | 45.24(1.36) |
| s5(t) | | | -12.55(0.99) | -12.21(0.26) | 11.67(0.49) |
| c6(t) | | | 71.46(6.80) | 61.79(1.50) | 66.83(3.25) |
| s6(t) | | | -4.32(0.31) | 6.38(0.12) | 1.63(0.06) |
| c7(t) | | | 9.39(1.34) | 7.29(0.30) | 8.19(0.66) |
| s7(t) | | | -22.19(3.15) | -22.71(0.93) | -23.47(-1.68) |
| c8(t) | | | 17.92(2.48) | 14.71(0.73) | 15.73(1.20) |
| s8(t) | | | 43.20(6.42) | 41.73(2.06) | 41.68(3.29) |

Table 8: Summary of discrete component models fitted to daylength series with (SE) or (t value) for estimates.

Note that using exact likelihood, in columns 2 and 4, the root of the AR operator rapidly moves away from unity. There are also some extremely large t values, although the estimates are not particularly stable due to the presence of the trend term with which they are correlated. Moving to the marginal likelihood estimates, the AR parameters remain close to those for the univariate model, in columns 3 and 5, and the t values are small. In the last column however the picture is quite different, with the roots of $\phi(B)$ now away from unity and moderate t values. Note however that the Mev values in columns 5 and 6, which are for separate local minima for the same model, have Mev values within one percent of each other. The marginal likelihood is designed for inference about the ARMA parameters, so we investigate further the structure of the Mev surface as a function of the two most variable parameters $\phi_1$ and $\phi_2$. The values of $\theta_1$ and $\theta_2$ which change very little for all the models are kept fixed at (-.537,-.615).

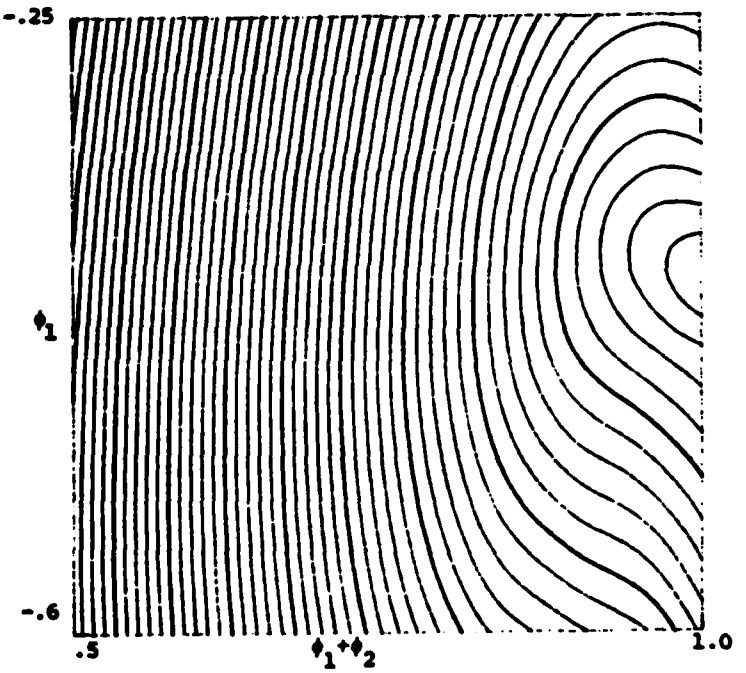



Figure 8: Contours of marginal likelihood of AR parameters in model for daylength variations with four frequency components.

Figures 8 and 9 show contour plots of this surface with $\phi_1$ on the vertical axis and $\phi_1 + \phi_2$ on the horizontal axis, so the right hand edge represents the boundary of the parameter space corresponding to a root of $\phi(B)$ at $B = 1$. Figure 8 is the map of Mev for 4 pairs of discrete components in the model. The minimum is on the boundary with no hint of any second local minimum. The classical confidence region would be given by the second heavy contour above the minimum, except that working close to the boundary the regularity conditions for properties of likelihood estimates will not hold. From a Bayesian point of view we do however have a reasonable approximation to a truncated bivariate Normal for which such a contour has an acceptable interpretation.

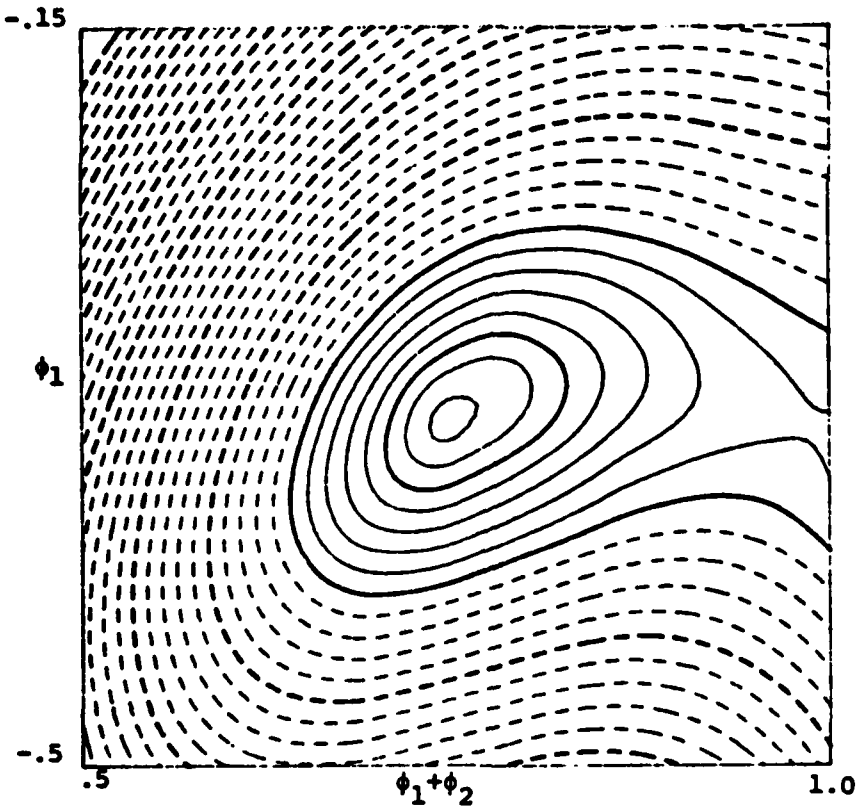



Figure 9: Contours of marginal likelihood of AR parameters, in model for daylength variations with eight frequency components.

Figure 9 is the map of Mev for 8 pairs of discrete components, and we see a quite different picture. Besides the obvious local minimum there is one on the boundary at $\phi_1 \simeq -.35$ and $\phi_1 + \phi_2 = 1$. The region between these is however very flat, the contours have been taken at fine intervals of Mev to reveal the structure. In this case the continuous contours cover a region where Mev is within one percent of its minimum, the broken contours being outside this. One interpretation might be to say that the scientist has to provide convincing evidence that his data cannot be easily explained by some random mechanism - in this case a random walk. Then there is very little evidence here to persuade one to forego the assumption that the parameters lie on the boundary minimum. The conclusion regarding the discrete components then comes from column 5 of Table 8 - none are convincingly significant. This may be viewed as being rather tough on the scientist who might claim that the approach used here could reduce any group of discrete spectrum components to insignificance, by choice of a suitable ARMA error structure. There is not much truth in this. Our contour maps are invariant to the magnitude of the discrete components in the model, so suitably large magnitudes could have lead to significant coefficients in column 5 of Table 8. It is possible for a series truly to consist of discrete components in such a mixture that they add up to look just like a random walk, but that would be just very bad luck for the scientist.

It is disconcerting, but not totally unexpected, that the exact likelihood ratio is so misleading in this case. Comparing the Dev value of 37990.88 for the model 1 in Table 5 with the Dev value of 21217.48 for the 8 component model in column 4 of Table 8, a mean deviance ratio of 101.19 would be referred to chi-square on 17 d.f. and taken as strong evidence of significant discrete components. To check that this was not due simply to the

selection of the frequencies by inspection of the sample spectrum peaks, the data was refitted using the harmonic frequencies corresponding to the periods in Table 7. An even smaller Dev value of 19691.95 resulted with very similar ARMA parameter values. As a further check a series of length 150 was simulated following the model 2 as shown in Table 4, i.e. with a random walk component. This is shown in Figure 10 to be broadly similar to the daylength series. After fitting the 8 component model by exact likelihood, a picture very close to that in column 4 of Table 8 emerged, with t values over 10.0. The harmonic frequencies were used here, and a mean deviance ratio of 81.8 resulted. Using Marginal likelihood, the Mev surface was again very flat over a neighbourhood of the boundary $\phi_1 + \phi_2 = 1$, with one local minimum on the boundary, slightly higher than a second minimum a short way off the boundary.

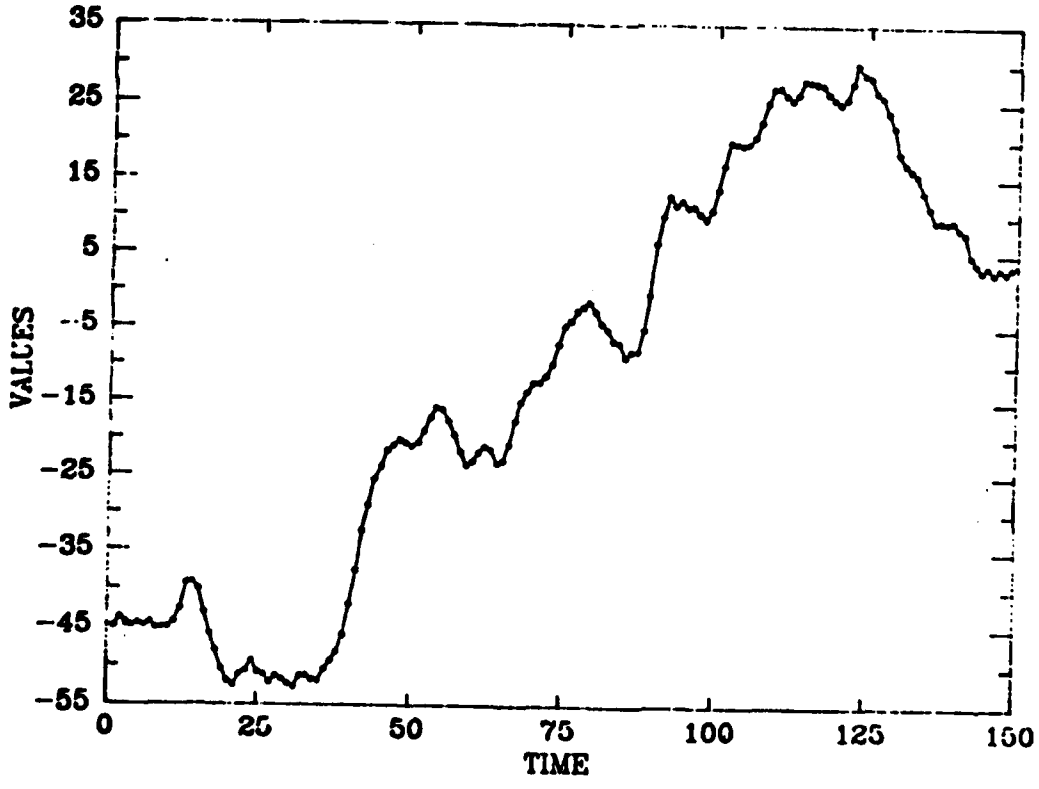



Figure 10: Simulated series from model 2 in table 5.

The grouping of a large number of discrete components at the end of the frequency range is in large part the cause of the difficulty. If for example the discrete components 6, 7 and 8 alone are fitted to the simulated data using exact likelihood, reasonable results are found, with a mean deviance ratio of 8.17 being referred to chi-square on 7 d.f.

A Bayesian approach may be applied in this case as for the regression parameter in the peak sunspot series of section 7. It is clear then that the high 't' values will be lost, and the low 't' values will tend to dominate. Taking a crude 50:50 mix at the two local minima in Figure 9 to approximate the posterior density of $(\phi_1,\phi_2)$, we see for example that for a typical regression coefficient with $\hat{\alpha} > 0$, $P(\alpha<0|y) > 1/2\ P(\alpha<0|y,\beta = \hat{\beta}_1)$ where $\hat{\beta}_1$ is the boundary point estimate for which the t value is small and the probability moderately large.

## 10. CONCLUSION

We have sought to demonstrate that marginal likelihood is a useful tool for dealing with a very practical problem - the estimation of the error structure in regression models. For those not satisfied with the Bayesian interpretation, there remains of course the question of the sampling properties of the estimates obtained using this criterion. There are other effects of using marginal likelihood, in particular the residual auto-correlations may not (and perhaps should not) look unduly 'white', since the estimation does not attempt to repair any distortion of the spectrum due to the regression.

In only one example was a non-deterministic regressor used (the peak sunspot numbers). This is because in time series applications, distributed lag

or transfer function models are often appropriate when true observed regressors are used. For such models with non-linear regression parameters, it is not immediately clear how to extend the marginal likelihood. The Bayesian approach is always possible provided suitable priors can be agreed. Local linearization might be a good initial step, and leads to formulae such as proposed by Leonard (1982) to approximate the results of integrating out nuisance parameters.

The software for marginal likelihood estimation used for the examples in this paper is available in the Genstat package and subroutine library distributed by the Numerical Algorithms Group, Oxford.

Although the use of marginal likelihood or its Bayesian equivalent has been advocated for many years, in the papers of Zellner and Tiao, Levenbach, and Cooper and Thompson, it has been somewhat neglected. It is hoped that this paper will help to stimulate more interest.

## REFERENCES


Abraham, B. and Box, G. E. P. (1978). Deterministic and Forecast-Adaptive Time-Dependent Models. Applied Statistics, 27, 120-130.

Ansley, C. F. and Newbold, P. (1980). Finite sample properties of estimators for autoregressive moving average models. Jour. Econometrics 13, 159-183.

Box, G. E. P. and Jenkins, G. M. (1976). Time Series Analysis, Forecasting and Control, 2nd Edition, San Francisco, Holden-Day.

Box, G. E. P. and Newbold, P. (1971). Some comments on a paper of Coen, Gomme and Kendall. J. R. Statist. Soc. (A) 134, 229-240.

Box, G. E. P. and Tiao, G. C. (1975). Intervention analysis with applications to economic and environmental problems. J. Amer. Statist. Assoc. 70, 70-79.

Campbell, M. J. and Walker, A. M. (1977). A survey of statistical work on the Mackenzie river series of annual Canadian lynx trappings for the years 1821-1934 and a new analysis. J. R. Statist. Soc. A, 140, 411-431.

Cooper, D. M. and Thompson, R. (1977). A note on the estimation of parameters of the autoregressive-moving average process. Biometrika, 64, 625-6.

Cryer, J. D. and Ledolter, J. (1981). Small sample properties of the parameters of the maximum likelihood estimator in the first order moving average model. Biometrika 68, 691-4.

Dicke, R. H. (1978). Is there a chronometer hidden deep in the sun? Nature, 276, 676-680.

Durbin, J. and Watson, G. S. (1950). Testing for serial correlation in least squares regression I. Biometrika, 37, 409-428.

Durbin, J. and Watson, G. S. (1951). Testing for serial correlation in least squares regression II. Biometrika, 38, 159-178.

Harvey, A. C. (1980). On comparing regression models in levels and first differences. Int. Econ. Rev. 21, 707-720.

Kalbfleisch, J. D. and Sprott, D. A. (1970). Application of Likelihood methods to models involving large numbers of parameters. JRSS(B) 32, 175-194.

Kendall, M. G. and Stuart, A. (1968). The advanced theory of statistics, Vol. 3 (Second edition) Hafner, New York.

King, M. L. (1983). Testing for autocorrelation in linear regression models: A survey. To appear in M. L. King and D. E. A. Giles (eds.), Specification analysis in the linear model: Essays in honour of Donald Cochrane.

Leonard, T. (1982). Discussion of Lejeune, M. and Faulkenberry, G. D. A simple predictive density function. J. Amer. Statist. Assoc. 77, 654-657.

Levenbach, H. (1972). Estimation of autoregressive parameters from a marginal likelihood function. Biometrika, 59, 61-71.

Luo, S. S., Liang, S. G., Ye, S. H., Yan, S. Z. and Li, Y. X. (1977). Analysis of periodicity in the irregular rotation of the earth. Chinese Astronomy 1, 221-227.

Sargan, J. D. and Bhargava, A. (1983). Testing residuals from least squares regression for being generated by the Gaussian random walk. Econometrica 51, 153-174.

Zellner, A. and Tiao, G. C. (1964). Bayesian Analysis of the Regression Model with autocorrelated errors. J. Amer. Statist. Assoc. 59, 763-778.

GTW/jvs

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

| REPORT DOCUMENTATION PAGE | | READ INSTRUCTIONS BEFORE COMPLETING FORM |
|---|---|---|
| 1. REPORT NUMBER<br>#2539 | 2. GOVT ACCESSION NO.<br>AD-A132828 | 3. RECIPIENT'S CATALOG NUMBER |
| 4. TITLE (and Subtitle)<br>On the Use of Marginal Likelihood in Time Series Model Estimation | | 5. TYPE OF REPORT & PERIOD COVERED<br>Summary Report - no specific reporting period<br>6. PERFORMING ORG. REPORT NUMBER |
| 7. AUTHOR(s)<br>G. Tunnicliffe-Wilson | | 8. CONTRACT OR GRANT NUMBER(s)<br>DAAG29-80-C-0041 |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS<br>Mathematics Research Center, University of Wisconsin<br>610 Walnut Street<br>Madison, Wisconsin 53706 | | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS<br>Work Unit Number 4 - Statistics and Probability |
| 11. CONTROLLING OFFICE NAME AND ADDRESS<br>U. S. Army Research Office<br>P.O. Box 12211<br>Research Triangle Park, North Carolina 27709 | | 12. REPORT DATE<br>July 1983<br>13. NUMBER OF PAGES<br>38 |
| 14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) | | 15. SECURITY CLASS. (of this report)<br>UNCLASSIFIED<br>15a. DECLASSIFICATION/DOWNGRADING SCHEDULE |

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Marginal likelihood, time series estimation, Durbin-Watson test, serial correlation, mixed spectrum

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper is concerned with the estimation of regression models with errors which follow an Autoregressive Integrated Moving Average (ARIMA) process. The effect of the regression upon the ARIMA model parameter estimates is considered and marginal likelihood investigated as a means of overcoming some small sample bias. Examples illustrate the importance of this effect even in samples of moderate size. The consequences regarding inference for the regression coefficients are also discussed.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

END

FILMED

10-83

DTIC